Data Center Network Topologies: VL2 (Virtual Layer 2)

Hakim Weatherspoon
Assistant Professor, Dept of Computer Science
CS 5413: High Performance Systems and Networking
September 26, 2014

Slides used and adapted judiciously from COS-561, Advanced Computer Networks At Princeton University
Goals for Today

• VL2: a scalable and flexible data center network
Architecture of Data Center Networks (DCN)

Key:
- BR = L3 Border Router
- AR = L3 Access Router
- S = L2 Switch
- LB = Load Balancer
- A = Rack of Servers
Conventional DCN Problems

- Static network assignment
- Fragmentation of resource
- Poor server to server connectivity
- Traffics affects each other
- Poor reliability and utilization
Objectives:

• Uniform high capacity:
  – Maximum rate of server to server traffic flow should be limited only by capacity on network cards
  – Assigning servers to service should be independent of network topology

• Performance isolation:
  – Traffic of one service should not be affected by traffic of other services

• Layer-2 semantics:
  – Easily assign any server to any service
  – Configure server with whatever IP address the service expects
  – VM keeps the same IP address even after migration
Measurements and Implications of DCN

- **Data-Center traffic analysis:**
  - Traffic volume between servers to entering/leaving data center is 4:1
  - Demand for bandwidth between servers growing faster
  - Network is the bottleneck of computation

- **Flow distribution analysis:**
  - Majority of flows are small, biggest flow size is 100MB
  - The distribution of internal flows is simpler and more uniform
  - 50% times of 10 concurrent flows, 5% greater than 80 concurrent flows
• Traffic matrix analysis:
  – Poor summarizing of traffic patterns
  – Instability of traffic patterns

• Failure characteristics:
  – Pattern of networking equipment failures: 95% < 1min, 98% < 1hr, 99.6% < 1 day, 0.09% > 10 days
  – No obvious way to eliminate all failures from the top of the hierarchy
Virtual Layer 2 Switch (VL2)

• Design principle:
  – Randomizing to cope with volatility:
    • Using Valiant Load Balancing (VLB) to do destination independent traffic spreading across multiple intermediate nodes
  – Building on proven networking technology:
    • Using IP routing and forwarding technologies available in commodity switches
  – Separating names from locators:
    • Using directory system to maintain the mapping between names and locations
  – Embracing end systems:
    • A VL2 agent at each server
Virtual Layer 2 Switch (VL2)

The Illusion of a Huge L2 Switch

1. L2 semantics
2. Uniform high capacity
3. Performance isolation
<table>
<thead>
<tr>
<th>Objective</th>
<th>Approach</th>
<th>Solution</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Layer-2 semantics</td>
<td>Employ flat addressing</td>
<td>Name-location separation &amp; resolution service</td>
</tr>
<tr>
<td>2. Uniform high capacity between servers</td>
<td>Guarantee bandwidth for hose-model traffic</td>
<td>Flow-based random traffic indirection (Valiant LB)</td>
</tr>
<tr>
<td>3. Performance Isolation</td>
<td>Enforce hose model using existing mechanisms only</td>
<td>TCP</td>
</tr>
</tbody>
</table>

“Hose”: each node has ingress/egress bandwidth constraints
Name/Location Separation

Cope with host churns with very little overhead

- Allows to use low-cost switches
- Protects network and hosts from host-state churn
- Obviates host and switch reconfiguration

Servers use flat names
Clos Network Topology

Offer huge aggr capacity & multi paths at modest cost

<table>
<thead>
<tr>
<th>D (# of 10G ports)</th>
<th>Max DC size (# of Servers)</th>
</tr>
</thead>
<tbody>
<tr>
<td>48</td>
<td>11,520</td>
</tr>
<tr>
<td>96</td>
<td>46,080</td>
</tr>
<tr>
<td>144</td>
<td>103,680</td>
</tr>
</tbody>
</table>

VL2

D ports

20 Servers

20*(DK/4) Servers
Cope with arbitrary TMs with very little overhead

[ ECMP + IP Anycast ]
• Harness huge bisection bandwidth
• Obviate esoteric traffic engineering or optimization
• Ensure robustness to failures
• Work with switch mechanisms available today

Equal Cost Multi Path Forwarding
1. Must spread traffic
2. Must ensure dst independence
VL2 Directory System

Directory Servers

Agent

“Lookup”

Agent

“Update”

1. Lookup

2. Reply

1. Update

5. Ack

2. Reply

4. Ack

(6. Disseminate)

3. Replicate

Directory Servers

RSM

Servers

RSM

RSM

RSM
Evaluation

• Uniform high capacity:
  – All-to-all data shuffle stress test:
    • 75 servers, deliver 500MB

  • Maximal achievable goodput is 62.3
  • VL2 network efficiency as 58.8/62.3 = 94%
Evaluation

- Fairness:
  - 75 nodes
  - Real data center workload
  - Plot Jain’s fairness index for traffics to intermediate switches
Evaluation

- Performance isolation:
  - Two types of services:
    - Service one: 18 servers do single TCP transfer all the time
    - Service two: 19 servers starts a 8GB transfer over TCP every 2 seconds
Evaluation

• Convergence after link failures
  – 75 servers
  – All-to-all data shuffle
  – Disconnect links between intermediate and aggregation switches
• Studied the traffic pattern in a production data center and find the traffic patterns

• Design, build and deploy every component of VL2 in an 80 server testbed

• Apply VLB to randomly spreading traffics over multiple flows

• Using flat address to split IP addresses and server names
Critique

• The extra servers are needed to support the VL2 directory system:
  – Brings more cost on devices
  – Hard to be implemented for data centers with tens of thousands of servers.

• All links and switches are working all the times, not power efficient

• No evaluation of real time performance.
• Similar “virtual layer 2” abstraction
  – Flat end-point addresses
  – Indirection through intermediate node

• Enterprise networks (Seattle)
  – Hard to change hosts $\rightarrow$ directory on the switches
  – Sparse traffic patterns $\rightarrow$ effectiveness of caching
  – Predictable traffic patterns $\rightarrow$ no emphasis on TE

• Data center networks (VL2)
  – Easy to change hosts $\rightarrow$ move functionality to hosts
  – Dense traffic matrix $\rightarrow$ reduce dependency on caching
  – Unpredictable traffic patterns $\rightarrow$ ECMP and VLB for TE
Other Data Center Architectures

• **VL2: A Scalable and Flexible Data Center Network**
  - consolidate layer-2/layer-3 into a “virtual layer 2”
  - separating “naming” and “addressing”, also deal with dynamic load-balancing issues

• **A Scalable, Commodity Data Center Network Architecture**
  - a new Fat-tree “inter-connection” structure (topology) to increases “bi-section” bandwidth
    - needs “new” addressing, forwarding/routing

Other Approaches:

• **PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric**

• **BCube: A High-Performance, Server-centric Network Architecture for Modular Data Centers**
Ongoing Research
Research Questions

• What topology to use in data centers?
  – Reducing wiring complexity
  – Achieving high bisection bandwidth
  – Exploiting capabilities of optics and wireless

• Routing architecture?
  – Flat layer-2 network vs. hybrid switch/router
  – Flat vs. hierarchical addressing

• How to perform traffic engineering?
  – Over-engineering vs. adapting to load
  – Server selection, VM placement, or optimizing routing

• Virtualization of NICs, servers, switches, …
Research Questions

• Rethinking TCP congestion control?
  – Low propagation delay and high bandwidth
  – “Incast” problem leading to bursty packet loss

• Division of labor for TE, access control, ...
  – VM, hypervisor, ToR, and core switches/routers

• Reducing energy consumption
  – Better load balancing vs. selective shutting down

• Wide-area traffic engineering
  – Selecting the least-loaded or closest data center

• Security
  – Preventing information leakage and attacks
Before Next time

- **Project Progress**
  - Need to setup environment as soon as possible
  - And meet with groups, TA, and professor

- **Lab0b – Getting Started with Fractus**
  - Use Fractus instead of Red Cloud
    - Red Cloud instances will be terminated and state lost
  - Due Monday, Sept 29

- **Required review and reading for Friday, September 26**
  - [http://dl.acm.org/citation.cfm?id=319166](http://dl.acm.org/citation.cfm?id=319166)

- Check website for updated schedule